

Energy-Efficient Classification for Anomaly Detection: The Wireless Channel as a Helper

Kiril Ralinovski, Mario Goldenbaum, and Sławomir Stańczak

Abstract

Anomaly detection has various applications including condition monitoring and fault diagnosis. The objective is to sense the environment, learn the normal system state, and then periodically classify whether the instantaneous state deviates from the normal one or not. A flexible and cost-effective way of monitoring a system state is to use a wireless sensor network. In the traditional approach, the sensors encode their observations and transmit them to a fusion center by means of some interference avoiding channel access method. The fusion center then decodes all the data and classifies the corresponding system state. As this approach can be highly inefficient in terms of energy consumption, in this paper we propose a transmission scheme that exploits interference for carrying out the anomaly detection directly in the air. In other words, the wireless channel helps the fusion center to retrieve the sought classification outcome immediately from the channel output. To achieve this, the chosen learning model is linear support vector machines. After discussing the proposed scheme and proving its reliability, we present numerical examples demonstrating that the scheme reduces the energy consumption for anomaly detection by up to 53 % compared to a strategy that uses time division multiple-access.

I. INTRODUCTION

Classification for anomaly detection is a common objective in areas such as image recognition, condition monitoring, and fault diagnosis [1]. The corresponding procedure is typically divided into two phases. During the first phase, the system environment is observed over a certain period of time in order to learn what the normal system state is. In the second phase, the instantaneous system state is sensed and then classified according to some rule (i.e., classifier) derived from the first phase. If the learning model was appropriately chosen, the classifier output indicates whether the system is currently in normal or abnormal state, where the latter is what is called

K. Ralinovski and S. Stańczak are with the Network Information Theory Group, Technische Universität Berlin, 10587 Berlin, Germany. M. Goldenbaum is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA.

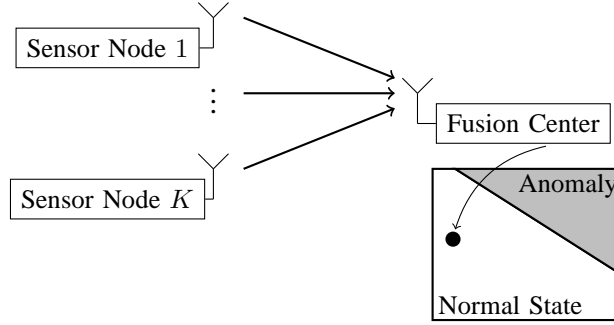


Fig. 1. In (wireless) anomaly detection, the objective is to classify an instantaneous system state based on the observations taken by a set of distributed sensor nodes. A conventional transmission strategy would provide that all the raw measurements are transferred to a fusion center where the classification is carried out.

an *anomaly* [2]. In industrial automation, for instance, the system state (or system condition) is monitored by means of spatially distributed temperature, oil pressure, and vibration sensors, just to name a few, that have a wired connection to some central processing unit.

A more flexible and cost-effective way to deal with it is to use a wireless sensor network (WSN) [3], [4]. A WSN, however, requires sensor nodes to be driven by batteries whose replacement is generally too expensive. Energy efficiency is therefore a major design objective in order to guarantee sufficiently long network lifetimes [5], [6]. A significant step in this direction is to drastically reduce the energy needed for communication. In a traditional WSN, the communication is as follows (see Fig. 1): the sensor nodes encode their observations and transmit them to a designated fusion center (FC) by means of some interference avoiding channel access method such as time division multiple-access (TDMA) or carrier sense multiple-access (CSMA). The FC then decodes all the data and classifies the instantaneous system state.

As this strategy can be highly inefficient in terms of energy consumption, in this paper we propose an anomaly detection scheme that is based on the communication strategy proposed and evaluated in [7]–[9]. The big difference to the traditional strategy outlined above is that *interference is exploited*, rather than avoided, for carrying out the anomaly detection directly *in the air*. In other words, the wireless channel helps the FC to retrieve the sought classification outcome immediately from the channel output. To achieve this, the learning model chosen in this paper is *linear support vector machines* [10, p. 112]. Due to its universality it is a widely used model and has recently been proven to be also a good candidate for condition monitoring

and fault diagnosis [1]. After adapting the transmission scheme of [7] to our specific needs, we prove that the resulting classification scheme is able to detect anomalies over the wireless channel with arbitrary small probability of error. In this context, we also discuss the trade-off between achievable accuracy and communication costs in terms of the number of transmissions needed. Finally, we present some numerical examples that demonstrate that the corresponding energy consumption can be reduced by up to 53 % compared to a strategy that uses TDMA.

A. Paper Organization

The rest of the paper is organized as follows. Section II provides the system model and the problem statement. In Section III, we summarize the basics of linear support vector machines as the chosen learning model. Subsequently, in Section IV we propose a transmission scheme that exploits interference for classifying system states and prove its performance. Section V provides some numerical examples and comparisons with a standard schemes, whereas Section VI concludes the paper.

B. Notational Remarks

Random variables are denoted by uppercase letters and their realizations by lowercase letters, respectively. Vectors are denoted by bold lowercase letters and matrices by bold uppercase letters. The Euclidean norm of some vector \mathbf{x} is denoted as $\|\mathbf{x}\|$. The expected value and the variance of some random element are denoted as $\mathbb{E}\{\cdot\}$ and $\mathbb{V}\text{ar}\{\cdot\}$, whereas $\mathbb{P}(A)$ designates the probability of any event A . Finally, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

II. SYSTEM MODEL AND PROBLEM STATEMENT

Consider a WSN that consists of a FC and a finite number of $K \geq 2$ spatially distributed nodes. As depicted in Fig. 1, the objective of the network is to monitor some system and periodically sample and classify its state in order reliably and efficiently detect anomalies. Towards this end, each node observes a coordinate of the corresponding K -dimensional system state $\mathbf{s}[t] := (S_1[t], \dots, S_K[t])^T$, $t \in \mathbb{N}$, assumed to be an element of the state space $\mathcal{S} := \mathcal{S}_1 \times \dots \times \mathcal{S}_K$. Here and hereafter, $\mathcal{S}_k \subset \mathbb{R}$, $k = 1, \dots, K$, represents the compact interval of values sensor k can

provide. We simply call this the *sensing range* of node k , which can be a range of temperatures, pressures, vibration intensities, etc.

We model the sensors' observations $\{S_k[t] \in \mathcal{S}_k\}_{t \in \mathbb{N}, k = 1, \dots, K}$, as time-discrete stochastic processes.¹ Each node is equipped with a transmitter in order to send its instantaneous observation to the FC. Thus, for every fixed t , each node, say node k , maps $S_k[t]$ to a length- M sequence of complex-valued transmit symbols $X_k^{(t)}[1], \dots, X_k^{(t)}[M]$, subject to some peak-power constraint $P_{\max} > 0$, that is, $\max_{1 \leq m \leq M} |X_k^{(t)}[m]|^2 \leq P_{\max}$. Allowing the nodes to transmit concurrently in the same frequency band, the sequence of symbols received by the FC can be modeled as [11]

$$Y^{(t)}[m] = \sum_{k=1}^K h_k X_k^{(t)}[m] + N^{(t)}[m], \quad m = 1, \dots, M. \quad (1)$$

Coefficient $h_k \in \mathbb{C}$ denotes the attenuation between node k and the FC, assumed to be known to node k , and N a proper complex Gaussian receiver noise process of variance σ_N^2 .

Assume that state space \mathcal{S} can be divided into classes “normal” and “anomaly” by a hyperplane. This means that at every time instant t , the system state $\mathbf{s}[t]$ belongs to either of these two classes. For ease of notation, we describe this by the set $\mathcal{C} := \{\pm 1\}$, where -1 represents the normal states and $+1$ anomalies. Then, the problem to be solved in this paper is to first learn from a subset of the sensors' observations a *classifier*

$$f : \mathcal{S} \rightarrow \mathcal{C}, \mathbf{s} \mapsto f(\mathbf{s}) \quad (2)$$

that is optimal in the sense that it minimizes false classifications. Given such classifier, we then wish to efficiently compute it over the wireless channel (1) such that

$$\forall \mathbf{s} \in \mathcal{S} : \lim_{M \rightarrow \infty} \mathbb{P}(f(\mathbf{s}) \neq \hat{f}(\mathbf{s})) = 0, \quad (3)$$

where \hat{f} denotes some estimator of f .

III. LINEAR SUPPORT VECTOR MACHINES

Learning the classifier (2) usually requires to choose an appropriate model. Due to its universality and simplicity, we therefore consider in the following the supervised learning model *linear support vector machines* [10]. To this end, we shall assume that we are provided with

¹This implies that system state $\mathbf{s}[t]$ is a time-discrete vector-valued stochastic process.

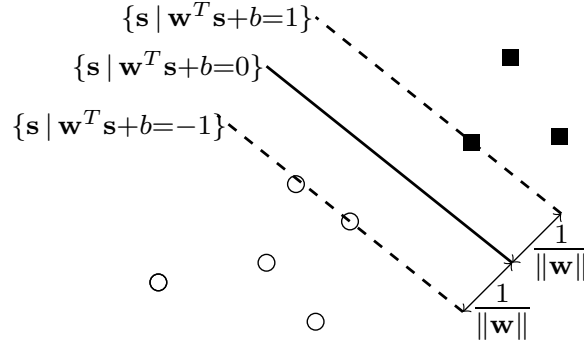


Fig. 2. A training set of system states that can be separated by a hyperplane. In this example, the *optimal* separating hyperplane is represented by the solid line. The dotted lines represent the boundary of the margin. State samples that lie on the boundary are called support vectors.

a *training set*, which is a sequence of already classified system states. To be more precise, for some $L \in \mathbb{N}$, the training set is defined as

$$\mathcal{T} := \{(s[l], c_l) \mid s[l] \in \mathcal{S}, c_l \in \mathcal{C}, l = 1, \dots, L\} \quad (4)$$

without loss of generality. Thus, for some realization of the system state, $s[l]$ denotes its l -th sample and c_l the class to which it belongs (i.e., -1 or 1). Given this training set, we are interested in determining an appropriate classifier f that is then used to reliably classify the remaining samples $s[L+1], s[L+2], \dots$ of the system state.

For linear support vector machines, (4) is supposed to be separable by a hyperplane

$$\mathbf{w}^T \mathbf{s} + b = 0, \quad \mathbf{s} \in \mathcal{S}, \quad (5)$$

determined by some nonvanishing normal vector $\mathbf{w} \in \mathbb{R}^K$ and some offset $b \in \mathbb{R}$. In this context, (5) is said to be *optimal* if it divides the training samples having $c_l = -1$ from those having $c_l = +1$ by a maximal margin [12].² Furthermore, (5) is said to be *canonical*, with respect to \mathcal{T} , if \mathbf{w} is normalized such that [12]

$$\min_{1 \leq l \leq L} |\mathbf{w}^T \mathbf{s}[l] + b| = 1.$$

See Fig. 2 for an illustration. Notice that if (5) is canonical, none of the training samples $s[1], \dots, s[L]$ lies inside the margin bounded by the two hyperplanes $\mathbf{w}^T \mathbf{s} + b = \pm 1$, $\mathbf{s} \in \mathcal{S}$.

²Note that maximizing the margin minimizes false classifications.

That is, for $l = 1, \dots, L$

$$\begin{aligned} \mathbf{w}^T \mathbf{s}[l] + b &\geq +1 & \text{if } c_l = +1 \\ \mathbf{w}^T \mathbf{s}[l] + b &\leq -1 & \text{if } c_l = -1, \end{aligned}$$

which can be summarized to

$$c_l(\mathbf{w}^T \mathbf{s}[l] + b) \geq 1, \quad l = 1, \dots, L. \quad (6)$$

Now, as the Euclidean distance between the boundaries (i.e., the width of the margin) follows to $\frac{2}{\|\mathbf{w}\|}$, finding the optimal separating hyperplane for the training set \mathcal{T} essentially requires to minimize the length of the normal vector \mathbf{w} :

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^K \setminus \{\mathbf{0}\}, b \in \mathbb{R}} \quad & \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & (6). \end{aligned} \quad (7)$$

The optimization problem (7) is a (convex) quadratic program [13] for which a variety of solvers exists (e.g., interior point methods [13, p. 561]). Denoting its optimal solution as (\mathbf{w}^*, b^*) , we finally conclude for the desired optimal classifier (2)

$$f(\mathbf{s}) = \text{sgn}(\mathbf{w}^{*T} \mathbf{s} + b^*), \quad (8)$$

where $\text{sgn} : \mathbb{R} \rightarrow \mathbb{R}$ denotes the classical signum function.

Remark 1. It should be emphasized that learning the optimal classifier as described in this section can be easily generalized to work with training sets that cannot be separated by a hyperplane. This is typically accomplished by adding slack variables in (7) [12, p. 131]. On the other hand, the method could also be modified to deal with unlabeled training sets [2].

IV. THE WIRELESS CHANNEL AS A HELPER

Now, as we know how the optimal classifier looks like, the question arises how to reliably and efficiently compute it over the channel (1). To solve this problem, most wireless system engineers would put effort in avoiding the interference in (1) by employing some channel access method such as TDMA or CSMA. It means that each node gets assigned its own time slot in which it transmits its instantaneous observation to the FC that subsequently classifies the system

state by evaluating (8). In this section, we go into the opposite direction and make explicit use of the interference by letting the nodes transmit simultaneously.

Towards this end, we adapt the transmission scheme of [7], which was designed to efficiently compute real-valued functions of distributed data over the wireless channel. As shown in [14], the scheme essentially allows to compute every function $f : \mathcal{S} \rightarrow \mathbb{R}$ that can be represented in the form

$$f(s_1, \dots, s_K) = \psi \left(\sum_{k=1}^K \varphi_k(s_k) \right) \quad (9)$$

by a proper choice of univariate functions $\varphi_k : \mathcal{S}_k \rightarrow \mathbb{R}$, $k = 1, \dots, K$, and $\psi : \mathbb{R} \rightarrow \mathbb{R}$. Functions of this kind are known as *nomographic functions* [15].

A closer look at (8) reveals that the optimal classifier is nomographic with $\varphi_k(s_k) = w_k^* s_k$ and $\psi(y) = \text{sgn}(y + b^*)$, where w_k^* and s_k denote the k -th elements of \mathbf{w}^* and \mathbf{s} , respectively. In the following two subsections, we describe how this special structure permits to let the wireless channel (1) help in the computation. In doing so, we assume that during network initialization, the quadratic program (7) was solved for some training set. The corresponding optimal coefficients w_k^* were then forwarded to the sensor nodes meaning that node k knows w_k^* prior to network operation. On the other hand, the FC was informed about b^* .

A. Transmitter

In what follows, without loss of generality we focus on an arbitrary time instant t and therefore drop the index for brevity. Each sensor node, say node k , encodes its *pre-processed* observation $\varphi_k(S_k) = w_k^* S_k$ as a *transmit power* into a random sequence of symbols. That is,

$$X_k[m] = \frac{1}{h_k} \sqrt{g(w_k^* S_k)} \exp(i\Theta_k[m]) , \quad (10)$$

for $m = 1, \dots, M$.³ The purpose of function

$$g : [\varphi_{\min}, \varphi_{\max}] \rightarrow [0, P_{\max}], g(\xi) = \frac{P_{\max}}{\varphi_{\max} - \varphi_{\min}} (\xi - \varphi_{\min})$$

is to ensure each node satisfies the transmit power constraint P_{\max} , where

$$\varphi_{\min} := \min_{1 \leq k \leq K} \min_{s \in \mathcal{S}_k} w_k^* s \quad (11a)$$

$$\varphi_{\max} := \max_{1 \leq k \leq K} \max_{s \in \mathcal{S}_k} w_k^* s . \quad (11b)$$

³Recall that $h_k \in \mathcal{C}$ was assumed to be a priori known to node k , $k = 1, \dots, K$ (cf. Section II). Thus, for $h_k \neq 0$, factor $1/h_k$ is for channel inversion. In fact, it has been shown in [16] that instead of h_k it even suffices to know its magnitude $|h_k|$.

In (10), the random phases $\{\Theta_k[m]\}_{k,m}$ are uniformly i.i.d. over $[0, 2\pi)$. Their task is to achieve a receiver-side decorrelation of the sensors' transmit signals, which simplifies the post-processing at the FC.

Remark 2. Encoding the pre-processed sensor observations as a transmit power has the advantage that quantization and precise symbol and phase synchronization is not needed. For further details, the reader is referred to [7].

B. Receiver

After receiving the M symbols in (1), the FC first determines the *receive energy* resulting in

$$\tilde{Y} := \sum_{m=1}^M |Y[m]|^2 = M \sum_{k=1}^K g(w_k^* S_k) + \tilde{N} . \quad (12)$$

It can be easily verified that the effective noise \tilde{N} has mean $\mathbb{E}\{\tilde{N}\} = M\sigma_N^2$ and variance

$$\begin{aligned} \text{Var}\{\tilde{N}\} &= 2M \sum_{k=1}^K \sum_{\substack{\ell=1 \\ \ell \neq k}}^K \mathbb{E}\{g(w_k^* S_k) g(w_\ell^* S_\ell)\} \\ &\quad + 2M\sigma_N^2 \sum_{k=1}^K \mathbb{E}\{g(w_k^* S_k)\} + M\sigma_N^4 . \end{aligned} \quad (13)$$

Except for g and \tilde{N} , the right-hand side of (12) is already equal to the inner product $\mathbf{w}^{*T} \mathbf{s}$. Thus, the FCs second signal post-processing step consists in applying the function

$$h : \mathbb{R} \rightarrow \mathbb{R} , \quad h(\tilde{y}) = \frac{\varphi_{\max} - \varphi_{\min}}{MP_{\max}} \tilde{y} + K\varphi_{\min} ,$$

(i.e., the counterpart to g), which results in

$$h(\tilde{Y}) = \mathbf{w}^{*T} \mathbf{s} + \frac{\varphi_{\max} - \varphi_{\min}}{P_{\max} M} \tilde{N} . \quad (14)$$

Let $\alpha := (\varphi_{\max} - \varphi_{\min})/P_{\max}$ and recall that \tilde{N} is not zero-mean. Then, in order to account for the bias induced by $\alpha M^{-1} \tilde{N}$, we finally define the estimate of (8) as

$$\hat{f}(\mathbf{s}) = \text{sgn}(h(\tilde{Y}) + b^* - \alpha M^{-1} \mathbb{E}\{\tilde{N}\}) , \quad (15)$$

which means that the FC finally adds b^* and $\alpha M^{-1} \mathbb{E}\{\tilde{N}\}$, followed by evaluating the sign of the resulting value.

C. Proving Reliability

Now, as we have introduced the transmission scheme for classifying system states by means of the wireless channel, in this subsection we prove its reliability in the sense that the corresponding probability of false classification can be made arbitrary small with increasing block length M .

Proposition 1. *Let f be the classifier given in (8) and \hat{f} its estimate as defined in (15). Then, with the transmission scheme proposed in Sections IV-A and IV-B, any system state $\mathbf{s} \in \mathcal{S}$ can be reliably classified, in the sense of (3), over the channel (1).*

Proof: Let $\mathbf{s} \in \mathcal{S}$ be arbitrary but fixed and note that

$$\mathbb{P}(\hat{f}(\mathbf{s}) \neq f(\mathbf{s})) = \mathbb{P}(|\hat{f}(\mathbf{s}) - f(\mathbf{s})| \geq 2)$$

as $f(\mathbf{s})$ and $\hat{f}(\mathbf{s})$ can only take on values $\{-1, 0, 1\}$.⁴ By virtue of Markov's inequality, it follows

$$\begin{aligned} \mathbb{P}(|\hat{f}(\mathbf{s}) - f(\mathbf{s})| \geq 2) &= \mathbb{P}((\hat{f}(\mathbf{s}) - f(\mathbf{s}))^2 \geq 4) \\ &\leq \frac{1}{4} \mathbb{E}\{(\hat{f}(\mathbf{s}) - f(\mathbf{s}))^2\}. \end{aligned} \quad (16)$$

Therefore, in order to prove that (3) applies it suffices to show that (16) vanishes with increasing block length M .

Towards this end, substitute (14) into (15), which yields

$$\hat{f}(\mathbf{s}) = \text{sgn} \left(\mathbf{w}^{*T} \mathbf{s} + b^* + \alpha M^{-1} (\tilde{N} - \mathbb{E}\{\tilde{N}\}) \right). \quad (17)$$

Let $\beta := \mathbf{w}^{*T} \mathbf{s} + b^*$ and $\Delta := \tilde{N} - \mathbb{E}\{\tilde{N}\}$. Then, since $\alpha, M > 0$, the signum function entails that

$$\hat{f}(\mathbf{s}) = \begin{cases} -1 & \text{if } \alpha\Delta/M < -\beta \\ 0 & \text{if } \alpha\Delta/M = -\beta \\ +1 & \text{if } \alpha\Delta/M > -\beta \end{cases}, \quad (18)$$

for all $\beta \in \mathbb{R}$. Bearing this in mind, we have to show that for each of the three cases $\beta > 0$, $\beta < 0$, and $\beta = 0$, (16) vanishes with increasing M .

We treat $\beta > 0$ first: In this case, the true classifier provides $f(\mathbf{s}) = f(\mathbf{s})^2 = 1$. Furthermore, for the event $\{\hat{f}(\mathbf{s}) = 0\}$ to occur, the point $\beta + \alpha\Delta/M$ has to lie on the separating hyperplane.

⁴For fixed $\mathbf{s} \in \mathcal{S}$, the randomness in $\hat{f}(\mathbf{s})$ only stems from the effective channel noise \tilde{N} .

Due to the distribution of $\alpha\Delta/M$, however, this event is of measure zero and therefore does not contribute to the expected value of $f(\mathbf{s})^2$. As a consequence, $\mathbb{E}\{\hat{f}(\mathbf{s})^2\}$ is equal to one and we obtain

$$\mathbb{E}\{(\hat{f}(\mathbf{s}) - f(\mathbf{s}))^2\} = 2(1 - \mathbb{E}\{\hat{f}(\mathbf{s})\}) . \quad (19)$$

Hence, to show that (16) tends to zero relaxes to show that $\mathbb{E}\{\hat{f}(\mathbf{s})\} \rightarrow 1$ for $M \rightarrow \infty$. We have

$$\begin{aligned} \mathbb{E}\{\hat{f}(\mathbf{s})\} &= 1 \cdot \mathbb{P}(\alpha\Delta/M > -\beta) + (-1) \cdot \mathbb{P}(\alpha\Delta/M < -\beta) \\ &= 2 \cdot \mathbb{P}(\alpha\Delta/M > -\beta) - 1 . \end{aligned}$$

Notice that for $M \rightarrow \infty$, $\alpha\Delta/M$ converges in distribution to a degenerated random variable of mean zero [7].⁵ As $\beta > 0$, this implies that $\mathbb{P}(\alpha\Delta/M > -\beta)$ goes to 1 and therefore $\lim_{M \rightarrow \infty} \mathbb{E}\{\hat{f}(\mathbf{s})\} = 1$ as desired.

To treat $\beta < 0$, we go along similar lines as before. In this case, however, the true classifier provides $f(\mathbf{s}) = -1$ so that (19) modifies to

$$\mathbb{E}\{(\hat{f}(\mathbf{s}) - f(\mathbf{s}))^2\} = 2(1 + \mathbb{E}\{\hat{f}(\mathbf{s})\}) .$$

For (16) to vanish, $\mathbb{E}\{\hat{f}(\mathbf{s})\}$ has to converge to -1 with increasing block length. It holds

$$\lim_{M \rightarrow \infty} \mathbb{E}\{\hat{f}(\mathbf{s})\} = 1 - 2 \lim_{M \rightarrow \infty} \mathbb{P}(\alpha\Delta/M < -\beta) . \quad (20)$$

Now, as β is strictly negative, the probability on the right-hand side of (20) tends to one and thus we have $\lim_{M \rightarrow \infty} \mathbb{E}\{\hat{f}(\mathbf{s})\} = -1$ as desired.

Finally, we have to treat $\beta = 0$, which means that $\mathbf{w}^{*T}\mathbf{s} + b^*$ lies on the separating hyperplane (see Fig. 2) and the true classifier provides $f(\mathbf{s}) = 0$. Note that from the perspective of anomaly detection, this is an unfavorable case as β does not belong to either of the two classes “normal” and “anomaly”. From the perspective of reliably computing f , however, we have to ensure that also in this case \hat{f} provides asymptotically accurate estimates. To see that this is indeed the case notice that for $\beta = 0$, (17) only depends on the sign of Δ . Therefore, as $\alpha\Delta/M$ converges in distribution to a degenerate random variable of mean zero, in the limit $M \rightarrow \infty$ we have $\hat{f}(\mathbf{s}) = 0$ with probability one.

As in each of the three cases $\mathbf{s} \in \mathcal{S}$ was chosen arbitrary, this concludes the proof. ■

⁵A scalar random variable is said to be degenerated if it has finite mean and zero variance. Due to (13) and the fact that α is finite, we have $\text{Var}\{\alpha\Delta/M\} = \alpha^2 \text{Var}\{\tilde{N}\}/M^2$, which tends to zero as $M \rightarrow \infty$.

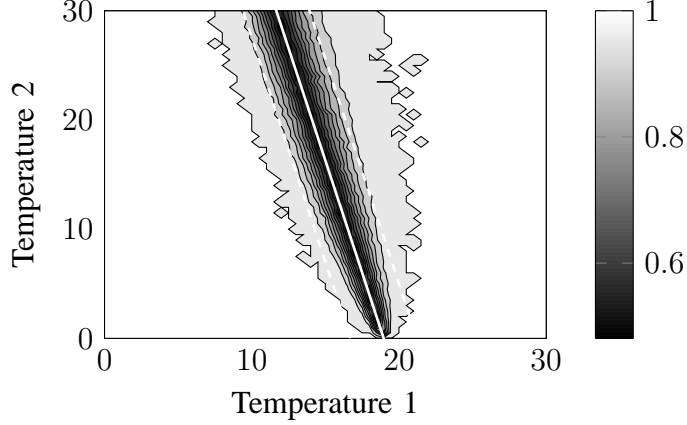


Fig. 3. Reliability map for $M = 64$ and $\text{SNR} = 10$ dB: the solid white line represents the separating hyperplane and the dashed lines the margin, respectively.

V. NUMERICAL EXAMPLES

In order to evaluate and discuss the performance of the proposed classification scheme in terms of reliability and energy efficiency, in this section we present some numerical examples. As the achievable performance depends for fixed M also on other system parameters such as transmit power constraint P_{\max} and noise variance σ_N^2 , we declare

$$\text{SNR} = \frac{\frac{1}{K} \sum_k \mathbb{E}\{P_k\}}{\sigma_N^2}, \quad (21)$$

with $P_k := g(w_k^* S_k)$ the instantaneous transmit power of node k , as the system operating point. Notice that (21) is the signal-to-noise ratio of a TDMA-based scheme, averaged over nodes.

A. Reliability

First, we present some plots that confirm the statement of Proposition 1. In addition, the plots reveal how many wireless transmissions are necessary in order to achieve a certain reliability.

Without loss of generality, let the number of sensors be $K = 2$, both of which are measuring a temperature in $^{\circ}\text{C}$. Let the corresponding sensing ranges be defined as $\mathcal{S}_1 = \mathcal{S}_2 = [0, 30]$, resulting in the state space $\mathcal{S} = [0, 30] \times [0, 30]$.

We assume that each instantaneous system state, $\mathbf{s} = (s_1, s_2)^T$, belongs equally likely to either of the two classes “normal” and “anomaly”, that is, $\mathbb{P}(\text{“normal”}) = \mathbb{P}(\text{“anomaly”}) = 0.5$. The

observations of class “normal” are distributed according to $(S_1, S_2) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and those of class “anomaly” as $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively, with⁶

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 20 \\ 20 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}.$$

In order to determine the optimal classifier (8), the support vector machines are trained with $L = 200$ samples drawn from these distributions. For block lengths $M = 2, \dots, 120$ and different signal-to-noise ratios, the reliability, measured as the conditional probability

$$\mathbb{P}\left(\hat{f}(S_1, S_2) = f(S_1, S_2) \mid (S_1, S_2) = (s_1, s_2)\right) \quad (22)$$

averaged over 10^4 Monte Carlo runs each, was then evaluated for 100 test samples.

Fig. 3 shows a reliability contour plot of the 100 test samples for $M = 64$ and SNR = 10 dB. It can be seen that the estimator (15) provides less reliable classifications when the system states are nearer to the separating hyperplane. In this specific example, the reliability significantly deteriorates within the margin, which is generally not the case as (22) is mainly determined by M and the SNR, whereas the margin is determined by the probability distribution of the system state.

The trade-off between the achievable reliability, block length M , and the SNR is depicted in Fig. 4. As it can be seen, the reliability monotonically improves as M increases. An interesting fact is that even at fairly low values of M and SNR, the proposed estimator \hat{f} provides results that are significantly above chance level 0.5 (i.e., above flipping a coin).

B. Energy Efficiency

Now, we analyze the efficiency of the proposed classification scheme in terms of energy consumption. In doing so, a scheme that uses a simplified TDMA acts as a reference in which each sensor node gets allocated $Q \in \mathbb{N}$ time slots to transmit its observation to the FC. In contrast to the preceding examples, we assume here that the network consists of $K = 32$ nodes and the state space is defined to be $S = [0, 30]^{32}$. Furthermore, the system states belonging to the classes “normal” and “anomaly” are distributed according to $\mathcal{N}(15 \cdot \mathbf{1}_K, 1.5 \cdot \mathbf{I}_K)$ and

⁶Strictly speaking, these distributions does not provide sensor values limited to the compact state space $[0, 30] \times [0, 30]$. However, with the explicit choice of covariance matrix $\boldsymbol{\Sigma}$, values outside this space are of negligible probability.

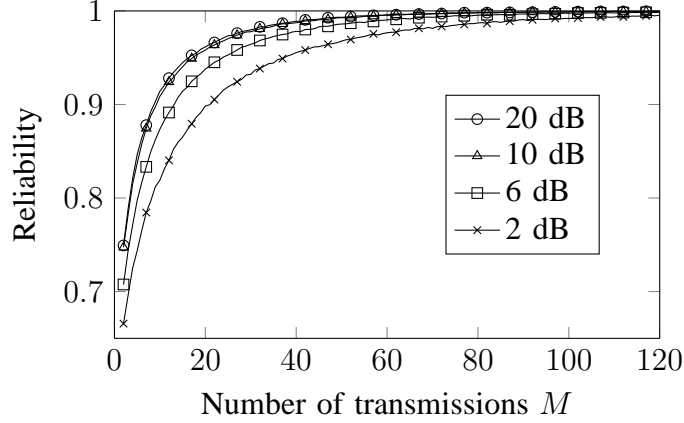


Fig. 4. Reliability as a function of the block length M and different signal-to-noise ratios.

$\mathcal{N}(20 \cdot \mathbf{1}_K, 1.5 \cdot \mathbf{I}_K)$, respectively, where $\mathbf{1}_K$ denotes the length- K vector of all ones and \mathbf{I}_K the $K \times K$ identity matrix. For each class, 1000 samples were generated. In what follows, “exploiting interference (EI)” refers to the proposed scheme.

To model the energy consumption, let $P_{\max} = 100 \mu\text{W}$ be the peak power and $T = 1 \text{ ms}$ the common symbol duration. These values are in compliance with the IEEE 802.15.4 standard [17]. Then, the transmit energies follow to $E_{\text{EI},k} = MP_kT$ and $E_{\text{TDMA},k} = QP_kT$, whereas the corresponding transmit durations are given by $T_{\text{EI}} = MT$ and $T_{\text{TDMA}} = QKT$, respectively. The overall energy consumption (per classification) is therefore given by $E_{\text{EI}} = \sum_{k=1}^K E_{\text{EI},k}$ and $E_{\text{TDMA}} = \sum_{k=1}^K E_{\text{TDMA},k}$. In order to compare the schemes in a fair manner, we set all values so that transmit durations and energy consumptions (per node) are equal (i.e., $T_{\text{EI}} = T_{\text{TDMA}}$ and $E_{\text{EI},k} = E_{\text{TDMA},k}$).

As can be seen from Fig. 5, the scheme proposed in this paper is much more energy efficient as the TDMA-base scheme. For instance, to achieve a reliability of 98% exploiting interference only requires 47% of the energy required with TDMA, due to the smaller number of transmissions required. In simple terms, the wireless channel helps to significantly improve the reliability as well as the energy efficiency.

VI. CONCLUSIONS

In this paper, the problem of reliably and efficiently classifying a system state for anomaly detection was considered. The problem occurs in many relevant areas such speech and image

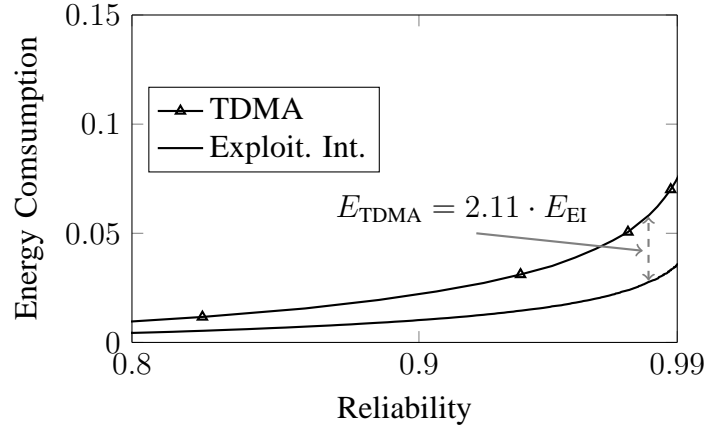


Fig. 5. Exploiting Interference (EI) vs. TDMA: energy consumption over reliability for $K = 32$ and $\text{SNR} = 1$ dB. At a reliability of 98 %, exploiting interference requires 53 % less energy than avoiding it.

recognition, condition monitoring, and fault diagnosis. To solve the problem, a novel classification scheme was proposed that is based on a set of wireless sensor nodes that monitor the system and simultaneously transmit their observations to a fusion center. In contrast to similar schemes known from the literature, the proposed scheme considers the wireless channel as helper. This means that the optimal classifier is essentially computed in the air by exploiting interference rather than avoiding it. The optimal classifier was learned from the sensor data based on the celebrated support vector machines model. Given this classifier, a corresponding estimator was proposed and shown to be able to make the probability of false classification arbitrary small. Finally, some numerical examples were presented that indicate that the proposed classification scheme is much more energy efficient as a scheme based on time-division multiple access. As a consequence, considering the wireless channel as a helper when transmitting the sensor observations to the fusion center may significantly increases the reliability as well as the energy efficiency.

REFERENCES

- [1] A. Widodo and B.-S. Yang, “Support vector machine in machine condition monitoring and fault diagnosis,” *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2560–2574, Aug. 2007.
- [2] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” in *Proc. 12th Neural Information Processing Systems Conf. (NIPS)*, Denver, CO, USA, Dec. 1999, pp. 582–588.

- [3] A. Willig, “Recent and emerging topics in wireless industrial communications: A selection,” *IEEE Trans. Ind. Informat.*, vol. 4, no. 2, pp. 102–124, May 2008.
- [4] H. Liqun and N. W. Bergmann, “Novel industrial wireless sensor networks for machine condition monitoring and fault diagnosis,” *IEEE Trans. Instrum. Meas.*, vol. 61, no. 10, pp. 2787–2798, Oct. 2012.
- [5] T. Rault, A. Bouabdallah, and Y. Challal, “Energy efficiency in wireless sensor networks: A top-down survey,” *Comput. Netw.*, vol. 67, no. 4, pp. 104–122, Jul. 2014.
- [6] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, Mar. 2002.
- [7] M. Goldenbaum and S. Stańczak, “Robust analog function computation via wireless multiple-access channels,” *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3863–3877, Sep. 2013.
- [8] A. Kortke, M. Goldenbaum, and S. Stańczak, “Analog computation over the wireless channel: A proof of concept,” in *Proc. IEEE Sensors 2014*, Valencia, Spain, Nov. 2014, pp. 1224–1227.
- [9] M. Goldenbaum, S. Stańczak, and H. Boche, “On achievable rates for analog computing real-valued functions over the wireless channel,” in *Proc. IEEE Int. Conf. Communications (ICC)*, London, UK, Jun. 2015, pp. 5639–5644.
- [10] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. New York: Cambridge University Press, 2000.
- [11] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. New York: Cambridge University Press, 2011.
- [12] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.
- [13] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.
- [14] M. Goldenbaum, H. Boche, and S. Stańczak, “Harnessing interference for analog function computation in wireless sensor networks,” *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4893–4906, Oct. 2013.
- [15] R. C. Buck, “Approximate complexity and functional representation,” *J. Math. Anal. Appl.*, vol. 70, no. 1, pp. 280–298, Jul. 1979.
- [16] M. Goldenbaum and S. Stańczak, “On the channel estimation effort for analog computation over wireless multiple-access channels,” *IEEE Wireless Commun. Lett.*, vol. 3, no. 3, pp. 261–264, Jun. 2014.
- [17] *IEEE Standard for Information Technology Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs)*, IEEE Std 802.15.4, 2003.